IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kilin Widjaja
25 March 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
  - Data Collection using Space X API and Web scraping
  - Data Wrangling
  - Exploratory Data Analysis using SQL and Visualization
  - Interactive Visual Analytics with Folium and Plotly Dash
  - Machine Learning Predictive Analysis
- Summary of all results
  - Exploratory Data Analytics shows which variables are best to predict success of landings
  - Interactive Visual Analytics and Dashboards help visualize the findings
  - Classification Predictive Analysis shows the result of the model used to predict success of landings

Introduction

- Project background and context

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, compare to other providers cost upward of 165 million dollars each. This much of savings is because SpaceX can reuse the first stage. This project is to help new company SpaceY to determine their cost of launches in order to compete with SpaceX.

- Problems you want to find answers

Therefore if we can determine if the first stage will land, we can determine the cost of a launch. In this project, we are predicting if the first stage will be able to land and reuse it.

Section 1

Methodology

Methodology

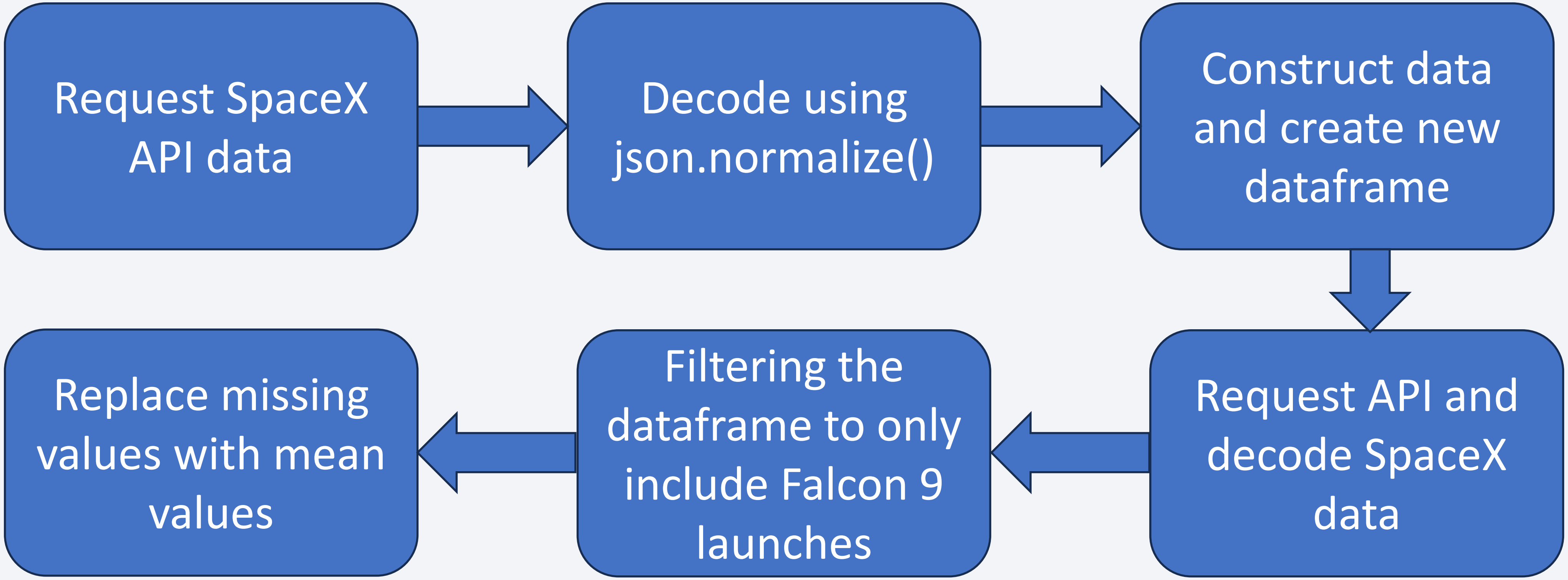
Executive Summary

- Data collection methodology
- Perform data wrangling
  - Converting the outcome into training labels with successful and unsuccessful outcome
- Perform exploratory data analysis (EDA) using Data Visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, evaluating classification models

Data Collection

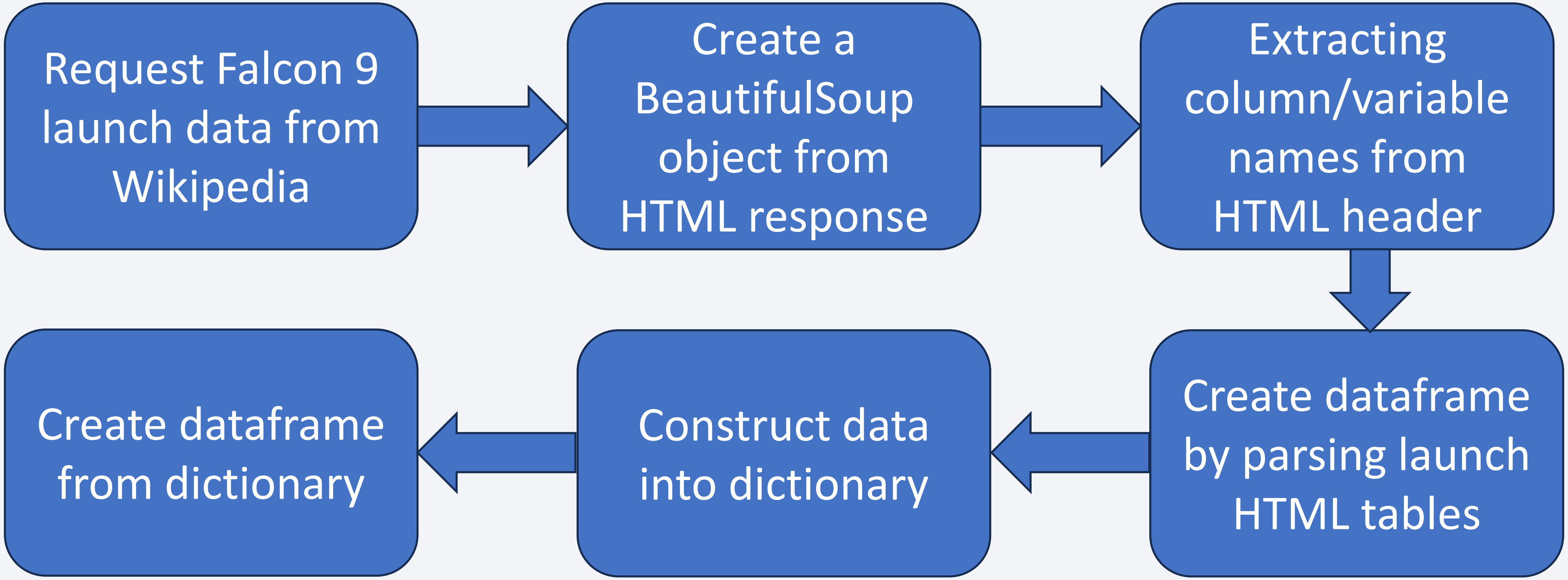
- Data sets were collected by using:
    - SpaceX API (https://api.spacexdata.com/v4/rockets/)
    - Web scraping from https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API



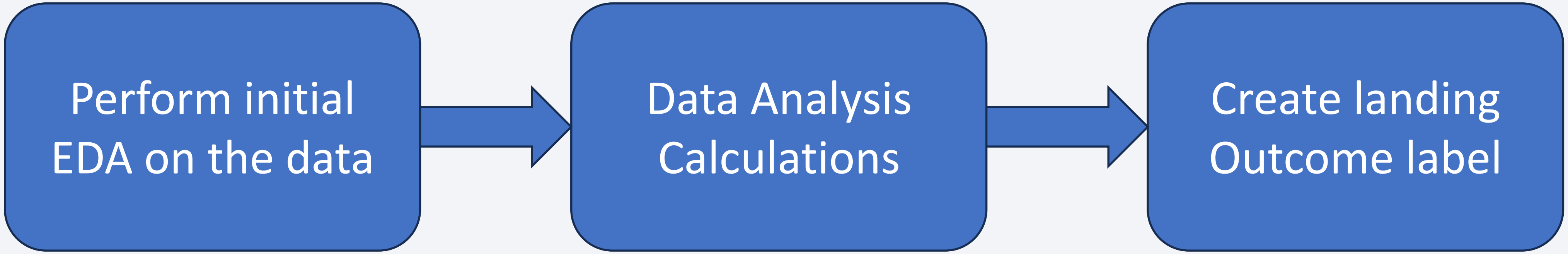
GitHub URL of the completed SpaceX API calls notebook, https://github.com/kilinwid/IBM-Applied-Data-Science-Capstone/blob/151f5277bf1243bd22ae8a3b627540477ac2f853/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



GitHub URL of the completed web scraping notebook, https://github.com/kilinwid/IBM-Applied-Data-Science-Capstone/blob/151f5277bf1243bd22ae8a3b627540477ac2f853/jupyter-labs-webscraping.ipynb

Data Wrangling



- Performing Data Analysis by:
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome of the orbits
  - Create a landing outcome label from Outcome column
- GitHub URL of the completed data wrangling notebook, https://github.com/kilinwid/IBM-Applied-Data-Science-Capstone/blob/151f5277bf1243bd22ae8a3b627540477ac2f853/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Plotted chart:
  - Flight Number vs. Pay Load Mass Chart
  - Flight Number vs. Launch Site Chart
  - Pay Load Mass vs. Launch Site Chart
  - Success Rate of each Orbit Type Chart
  - Flight Number vs. Orbit Type Chart
  - Pay Load Mass vs. Orbit Type Chart
  - Launch Success Yearly Trend
- These charts are plotted to see the relationship between two variables
- GitHub URL of the completed EDA with data visualization notebook, https://github.com/kilinwid/IBM-Applied-Data-Science-Capstone/blob/151f5277bf1243bd22ae8a3b627540477ac2f853/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- The SQL queries performed:
  - Display the names of unique launch sites in the space mission
  - Display top 5 launch sites that name begin with string ‘CCA’
  - Display total Payload Mass carried by boosters launched by NASA (CRS)
  - Display the average Payload Mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters which have success in drone ship and payload mass between 4000 and 6000 kg
  - List the total number of successful and failure mission outcomes
  - List the names of the booster version which have carried the maximum payload mass
  - List the failed landing outcomes in drone ship, booster version, and launch site for the months in year 2015
  - Rank the landing count outcomes between 2010-06-04 and 2017-03-20 in descending order
- GitHub URL of the completed EDA with SQL notebook, https://github.com/kilinwid/IBM-Applied-Data-Science-Capstone/blob/151f5277bf1243bd22ae8a3b627540477ac2f853/jupyter-labs-eda-sql-coursera_sqllite.ipynb

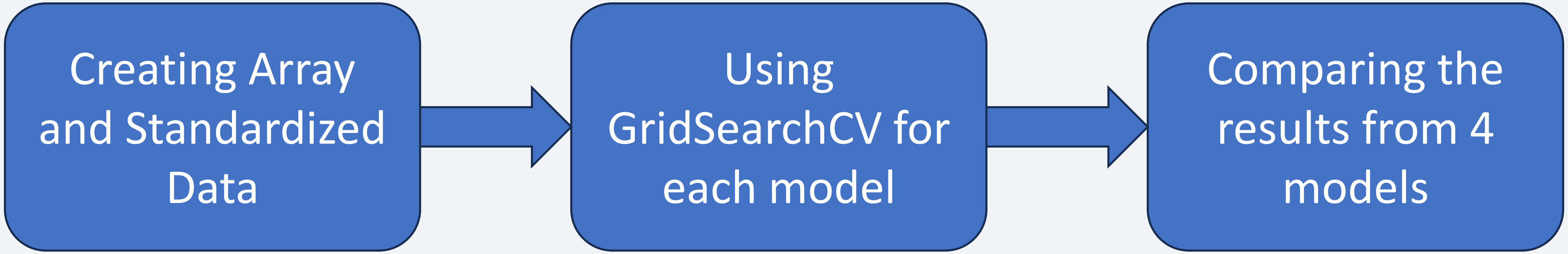
Build an Interactive Map with Folium

- Create Folium Map with:
  - Marking all launch sites on the map
  - Marking all successful or failed launches for each launch site on the map
  - Calculate the distance between from launch site to its proximities
  - Create line marker between launch site to its proximities
- The markers are used to help us visualize better about each launch sites on the map
- GitHub URL of the completed interactive map with Folium map, https://github.com/kilinwid/IBM-Applied-Data-Science-Capstone/blob/151f5277bf1243bd22ae8a3b627540477ac2f853/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Interactive Dashboard made by Plotly:
  - There are two part, first part is pie chart and the second part is scatter plot
  - Both chart and plot are Payload Mass vs Success Rate
  - Pie chart can be changed interactively by selecting Launch Site
  - Scatter plot can be change interactively by selecting Payload Mass
- This interactive dashboard can quickly visualize which launch site has better success rate considering its launch's payload mass
- GitHub URL of your completed Plotly Dash lab, https://github.com/kilinwid/IBM-Applied-Data-Science-Capstone/blob/151f5277bf1243bd22ae8a3b627540477ac2f853/spacex_dash_app.py

Predictive Analysis (Classification)



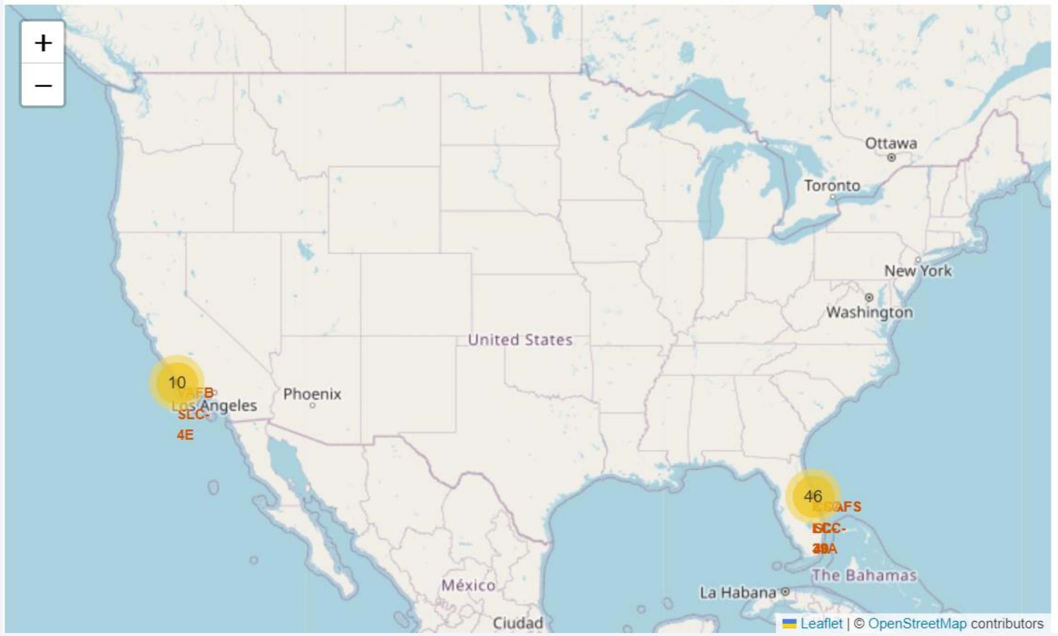
- For Machine Learning Predictive Analysis, we use four models: Logistic Regression, Support Vector Machine, Decision Tree, and K nearest neighbors
- GitHub URL of the completed predictive analysis lab, https://github.com/kilinwid/IBM-Applied-Data-Science-Capstone/blob/151f5277bf1243bd22ae8a3b627540477ac2f853/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

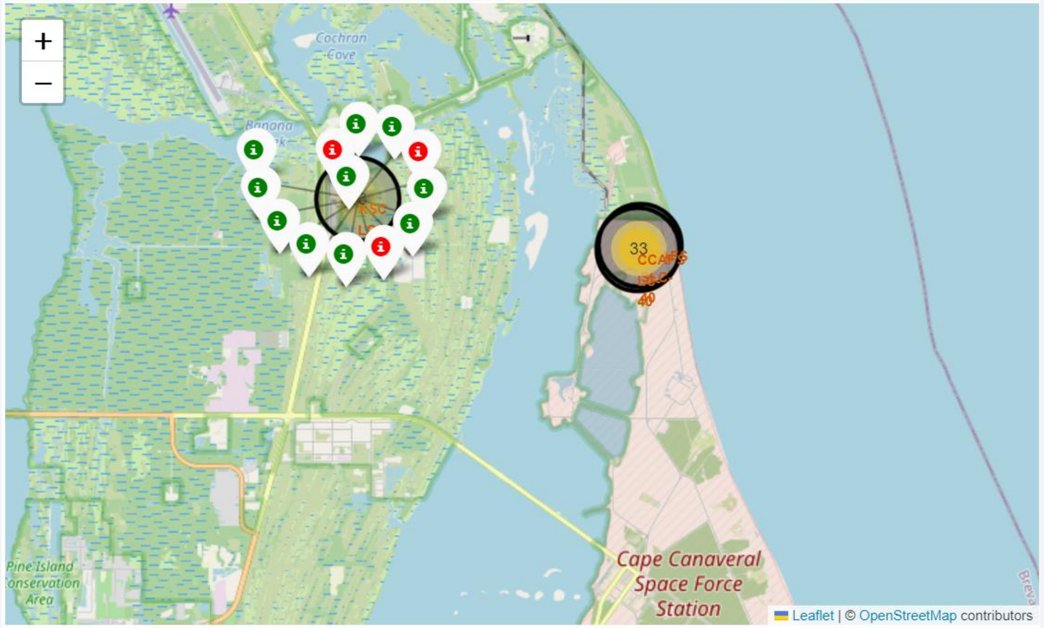
Exploratory Data Analysis Results

- There are four launch sites used by SpaceX
- The first successful landing on ground pad is on December 22, 2015
- There are four type of booster used for Payload Mass between 4000 and 6000 kg
- Only one failed mission outcome from 101 launches
- There are twelve type booster successfully launched maximum Payload Mass (15,600 kg)
- There are two type booster failed landing outcome in 2015
- The most successful landing outcome is using drone ship (5 times)
- Landing outcome success rate is continuously getting higher, starting from 2014

Interactive Analytics Demo Results

- There are 10 launches on west coast and 46 launches on east coast
- KSC LC-39A has 77% success rate, which is the highest among all four sites





Predictive Analysis Results

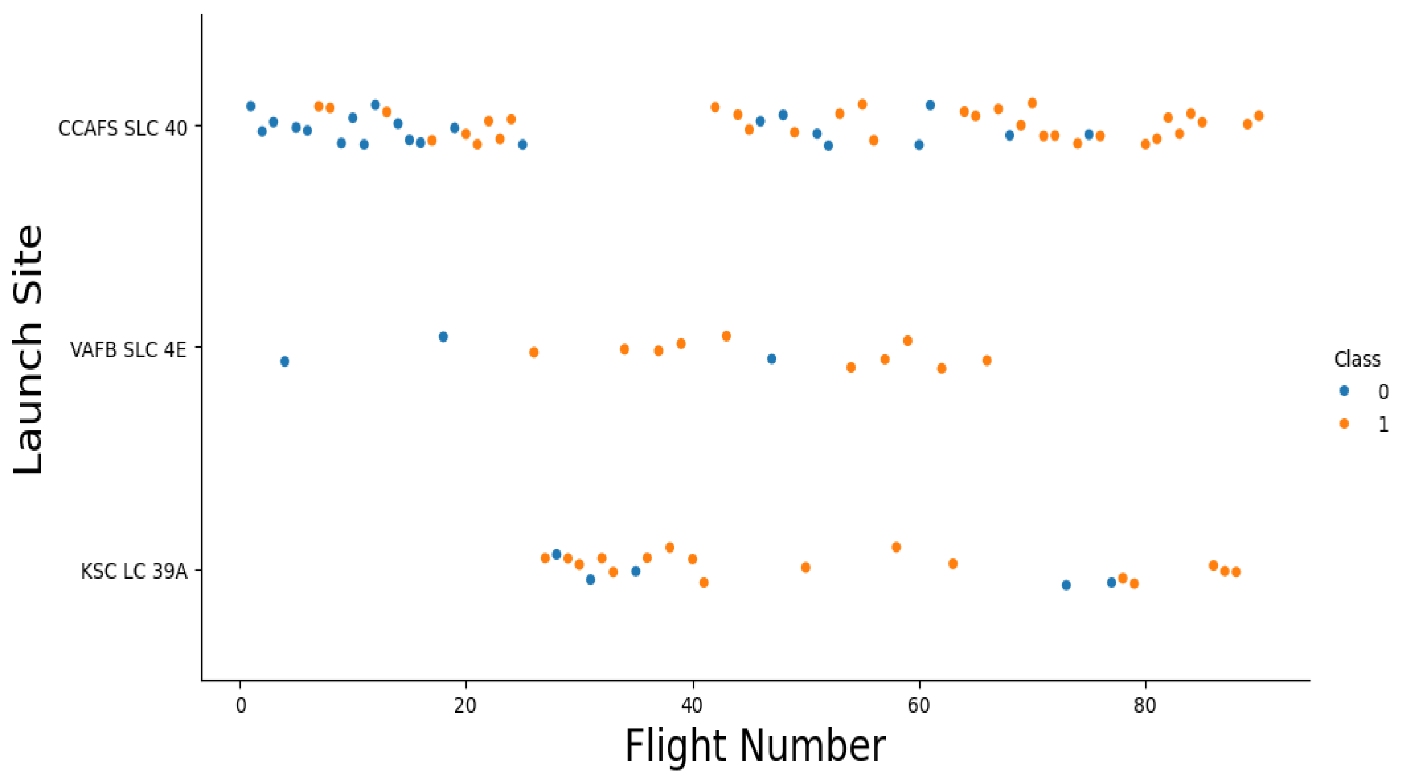
- Logistic Regression has an accuracy of 84.6% and test score of 83.3%
- Support Vector Machine has an accuracy 84.8% of and test score of 83.3%
- Decision Tree has an accuracy of 86.3% and test score of 83.3%
- K-nearest neighbors has an accuracy of 84.8% and test score of 83.3%
- We have the same test score for all four methods

Section 2

Insights drawn from EDA

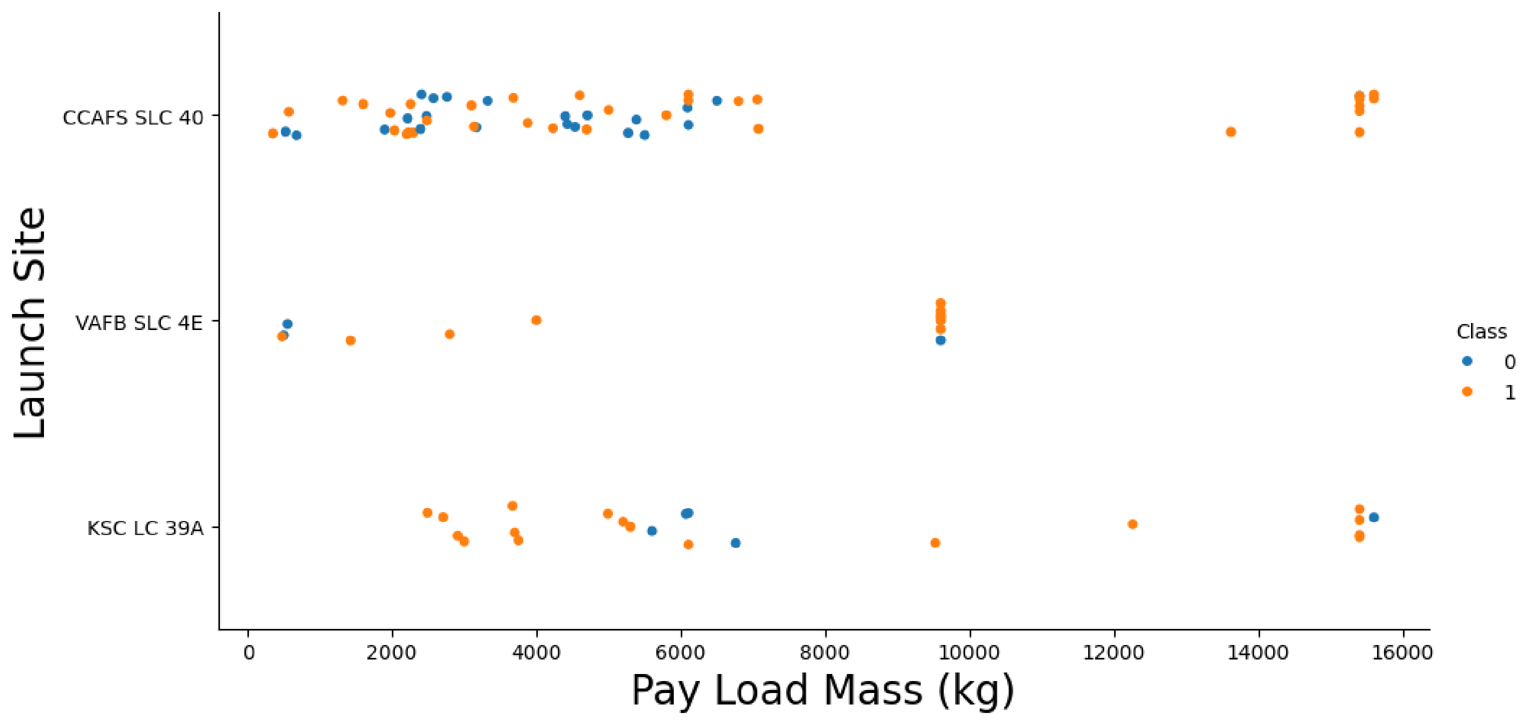
Flight Number vs. Launch Site

- According to the graph, general success rate has improved over time
- CCAFS SLC-40 has more launches than the other two sites
- In recent launches (after flight number 60), VAFB SLC-4E has the highest success (minimal launches), CCAFS SLC-40 come second (most launches)



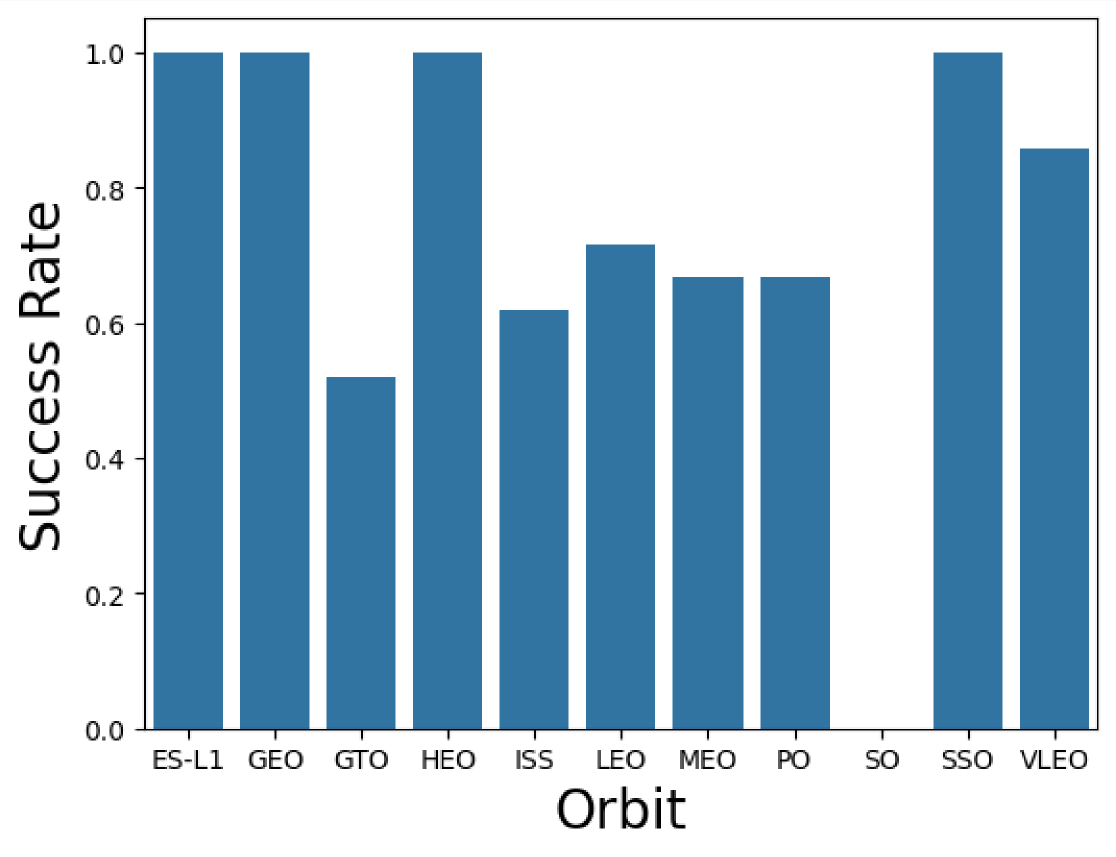
Payload vs. Launch Site

- Payload Mass over 8,000 kg has higher success rate
- Payload Mass over 12,000 kg can only launched from CCAFS SLC-40 and KSC LC-39A



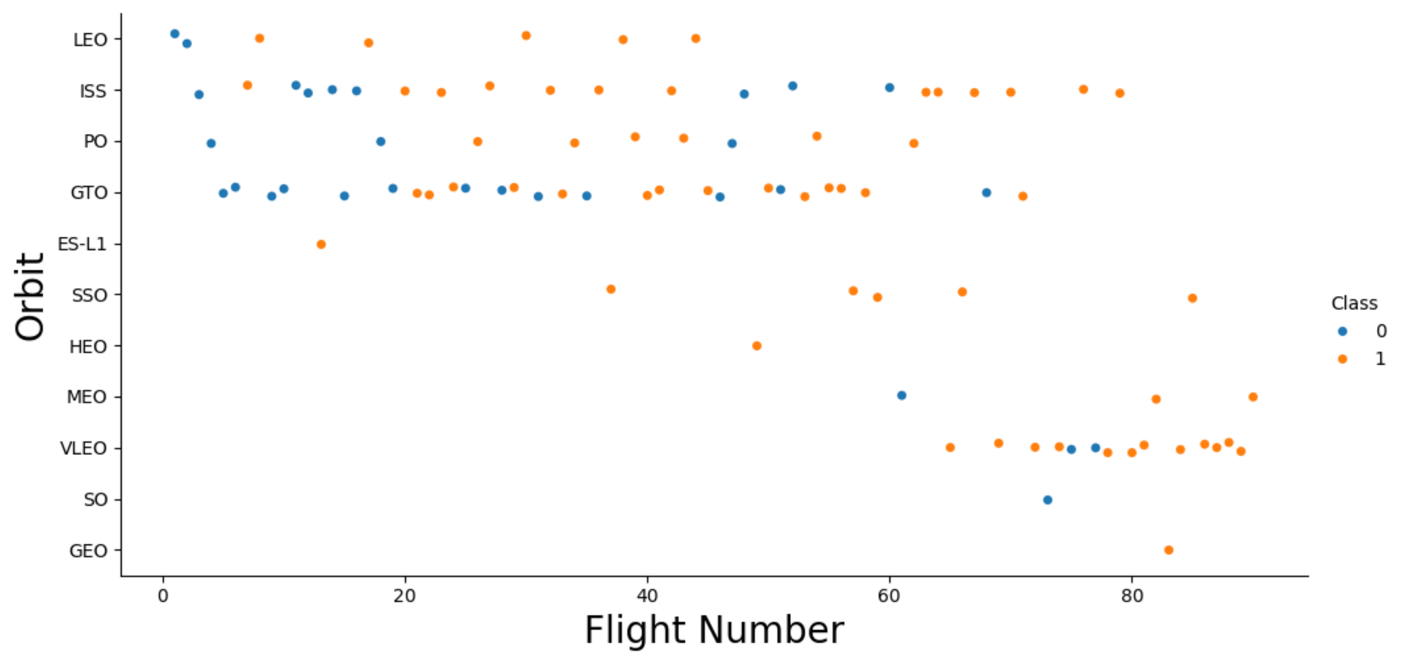
Success Rate vs. Orbit Type

- Four Orbits has 100% success rate: ES-L1, GEO, HEO, and SSO
- SO has 0% success rate
- The other orbits have success rate in between



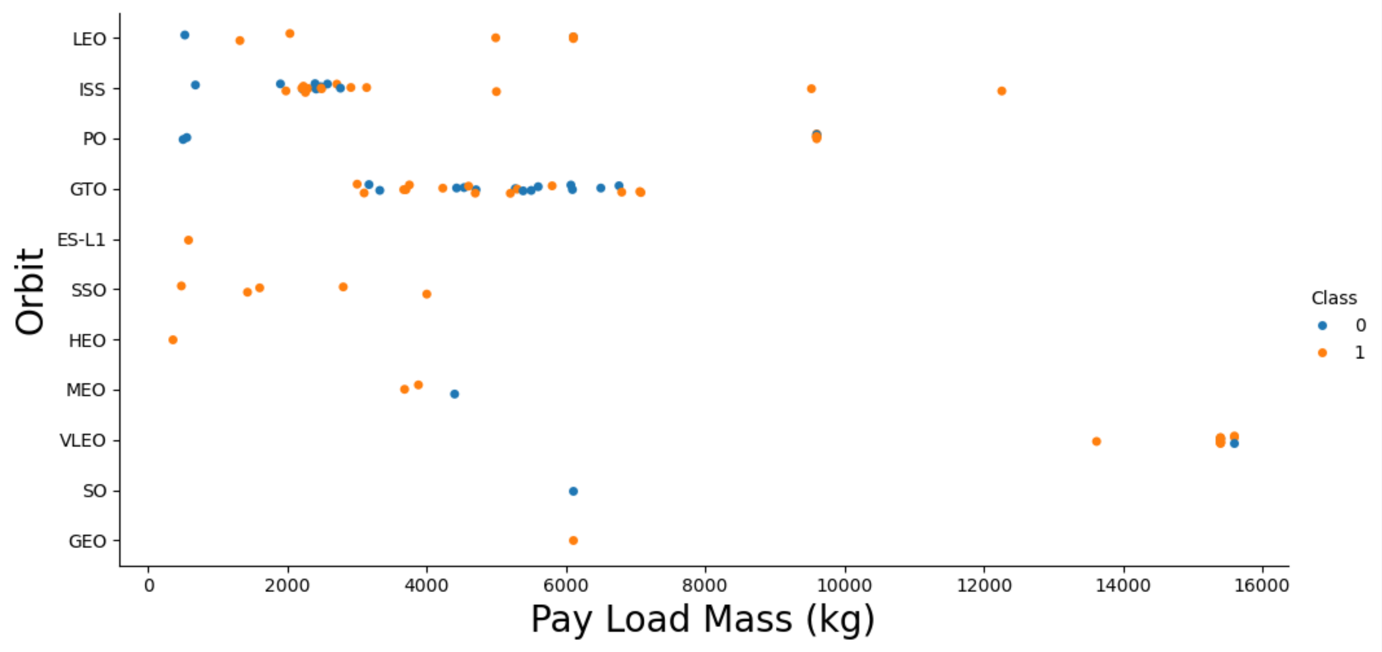
Flight Number vs. Orbit Type

- From the graph shown that success rate improve over time
- LEO has better success after first two fail rate
- GTO has some mix results



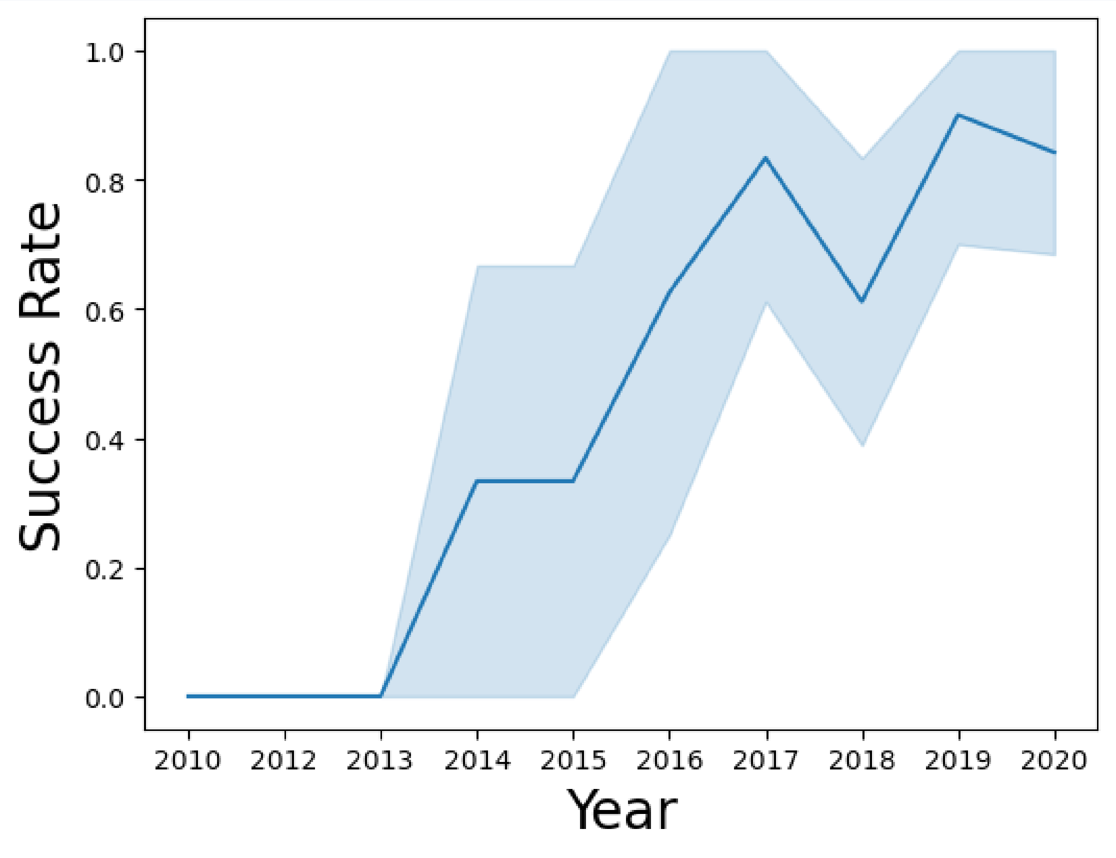
Payload vs. Orbit Type

- GTO and VLEO have mixed results
- LEO and ISS have better success rate over higher Payload Mass



Launch Success Yearly Trend

- From the graph shows that SpaceX has 3 year learning curve to adjust
- Success rate has improvement trend from 2014 to 2020



All Launch Site Names

- Using the query, there are four unique launch sites according to the data

Task 1

Display the names of the unique launch sites in the space mission ¶

```
[13]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
```

[13]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- Using the query for five sample launch data that begin with string ‘CCA’

Task 2

Display 5 records where launch sites begin with the string 'CCA'

[14]:
```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE '%CCA%' LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

[14]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Using the query to calculate the total payload carried by boosters from NASA

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

[20]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS "Payload Mass (KG)",Customer FROM SPACEXTABLE WHERE Customer='NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

[20]:

Payload Mass (KG)	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- Using the query to calculate the average payload mass carried by booster version F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

```
[28]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS "Average Payload Mass (KG)",Booster_Version FROM SPACEXTABLE WHERE Booster_Version='F9 v1.1';
```

```
 * sqlite:///my_data1.db
Done.
```

[28]:

Average Payload Mass (KG)	Booster_Version
2928.4	F9 v1.1

First Successful Ground Landing Date

- Using the query to find the dates of the first successful landing outcome on ground pad

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

[29]:
```
%sql SELECT MIN(DATE),Landing_Outcome FROM SPACEXTABLE WHERE Landing_Outcome='Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

[29]:

MIN(DATE)	Landing_Outcome
2015-12-22	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- Using the query to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 ¶

[30]:
```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome='Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

[30]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Using the query to calculate the total number of successful and failure mission outcomes

Task 7

List the total number of successful and failure mission outcomes

```
[34]: %sql SELECT Mission_Outcome, COUNT(Mission_Outcome) AS "TOTAL" FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

[34]:

Mission_Outcome	TOTAL
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Using the query to list the names of the booster which have carried the maximum payload mass

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

[35]:
```
%sql SELECT Booster_Version,PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

```
 * sqlite:///my_data1.db
Done.
```

[35]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- Using the query to list the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015. ¶

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

[38]:
```
%sql SELECT substr(Date, 6,2),substr(Date,0,5),Booster_Version,Launch_Site,Landing_Outcome FROM SPACEXTABLE WHERE Landing_Outcome='Failure (drone ship)' AND substr(Date,0,5)='2015';
```

```
 * sqlite:///my_data1.db
Done.
```

[38]:

substr(Date, 6,2)	substr(Date,0,5)	Booster_Version	Launch_Site	Landing_Outcome
01	2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Using the query to rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Task 10

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

[65]:
```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) AS "Total" FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY "Total" DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

[65]:

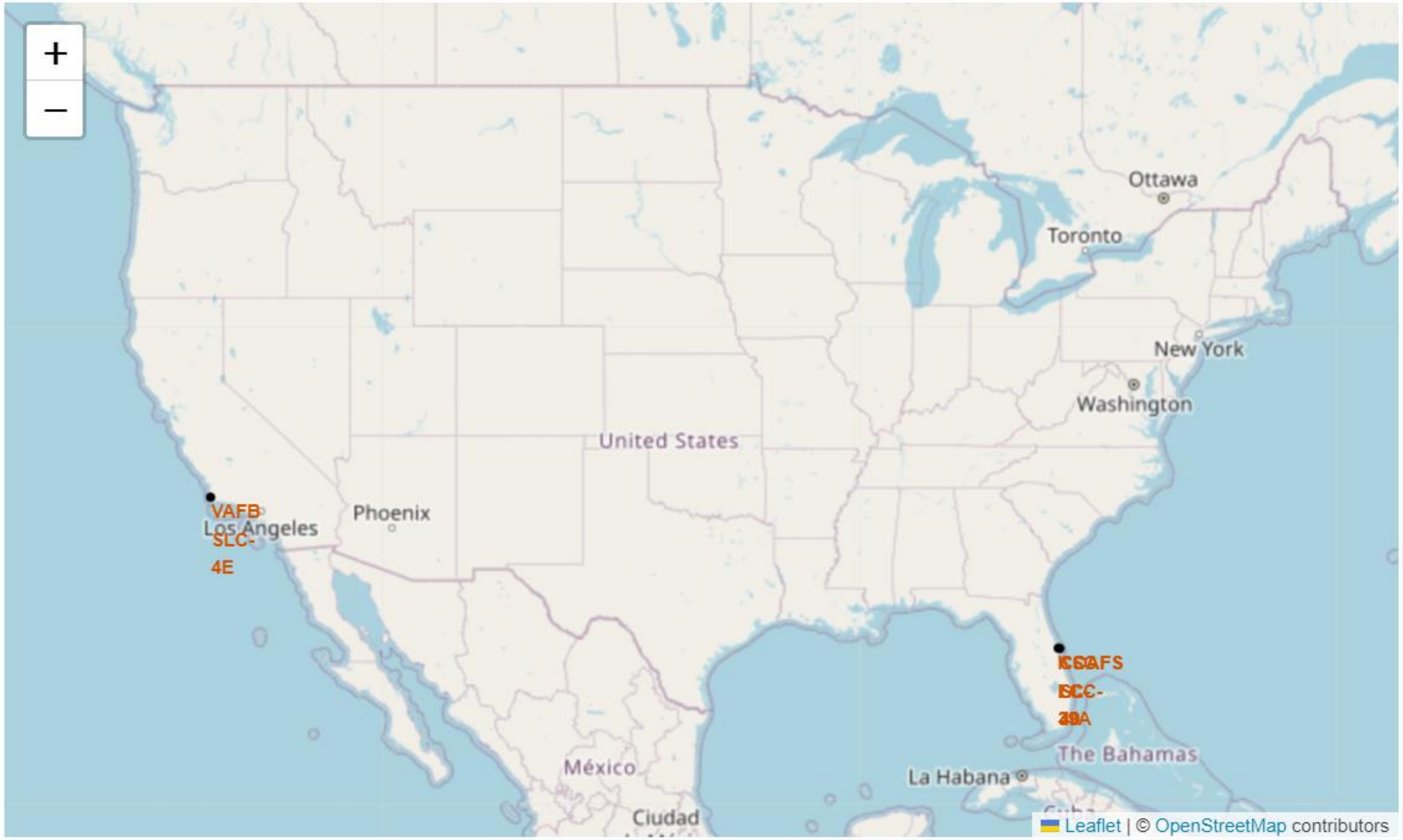
Landing_Outcome	Total
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

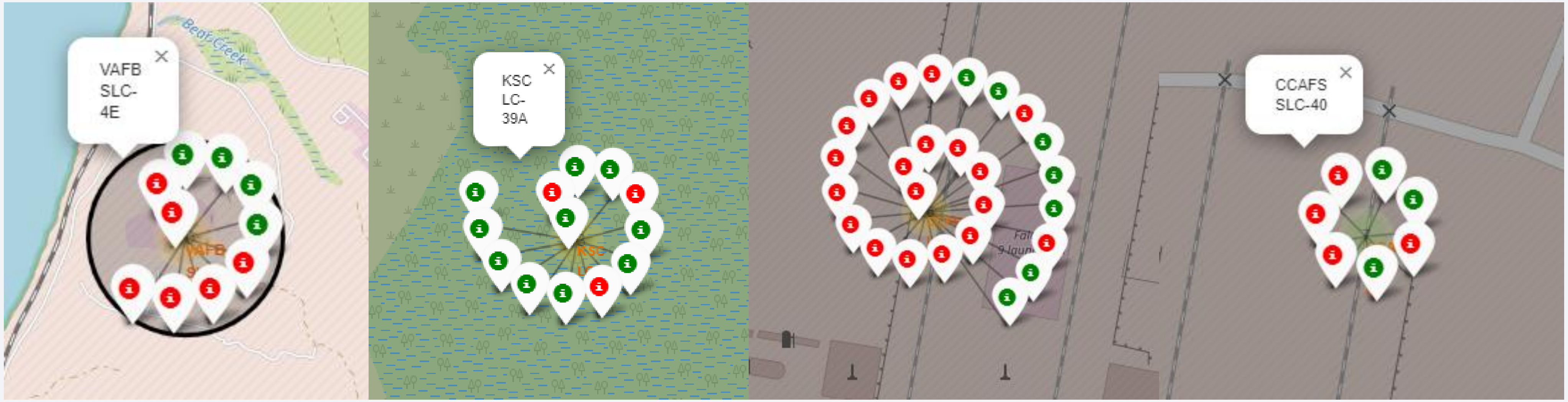
Launch Sites

- There is one launch site in West coast and three launch sites in East coast
- All four launch site is near coast line



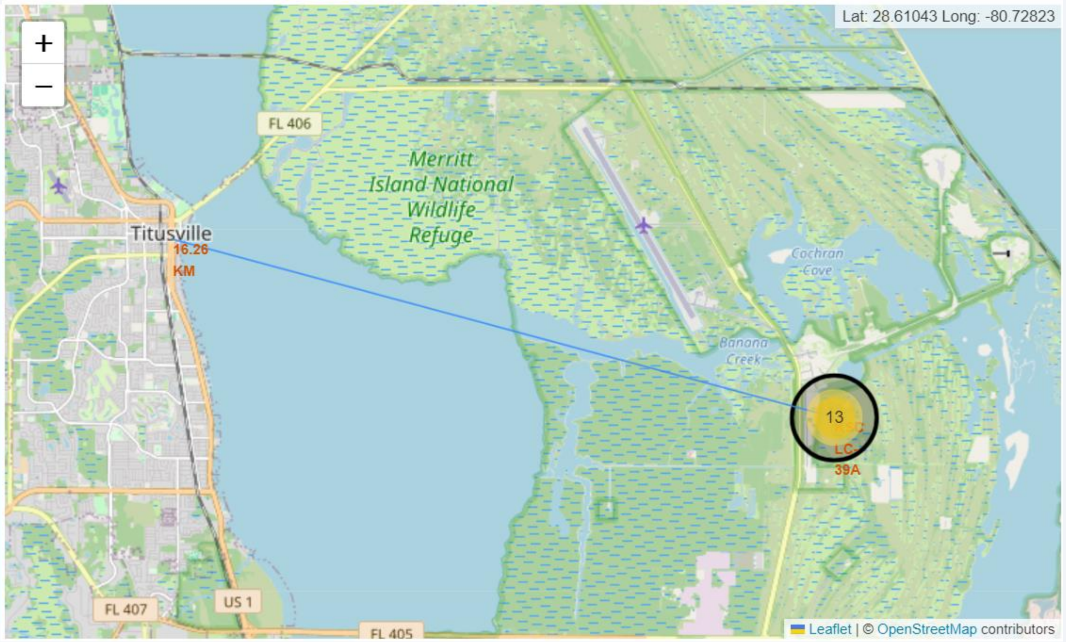
Site Launch Outcomes

- Below is the four launch sites outcomes, from left, VAFB SLC-4E, KSC LC-39A, CCAFS LC-40, CCAFS SLC-40
- Best outcome is from launch site KSC LC-39A



Distance from Launch Site to its proximities

- Distance to coast line is 0.86 km
- Distance to Titusville (closest city) is 16.26 km

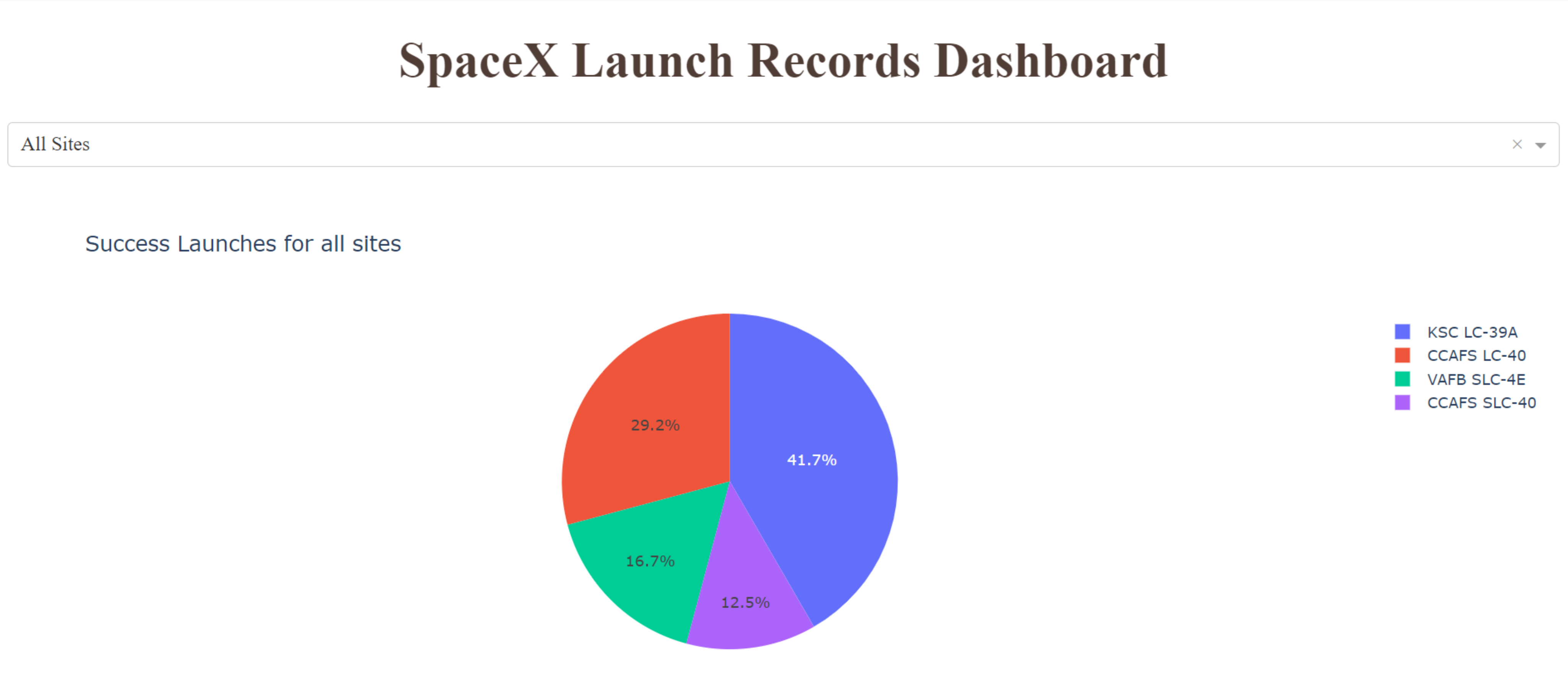


Section 4

Build a Dashboard with Plotly Dash

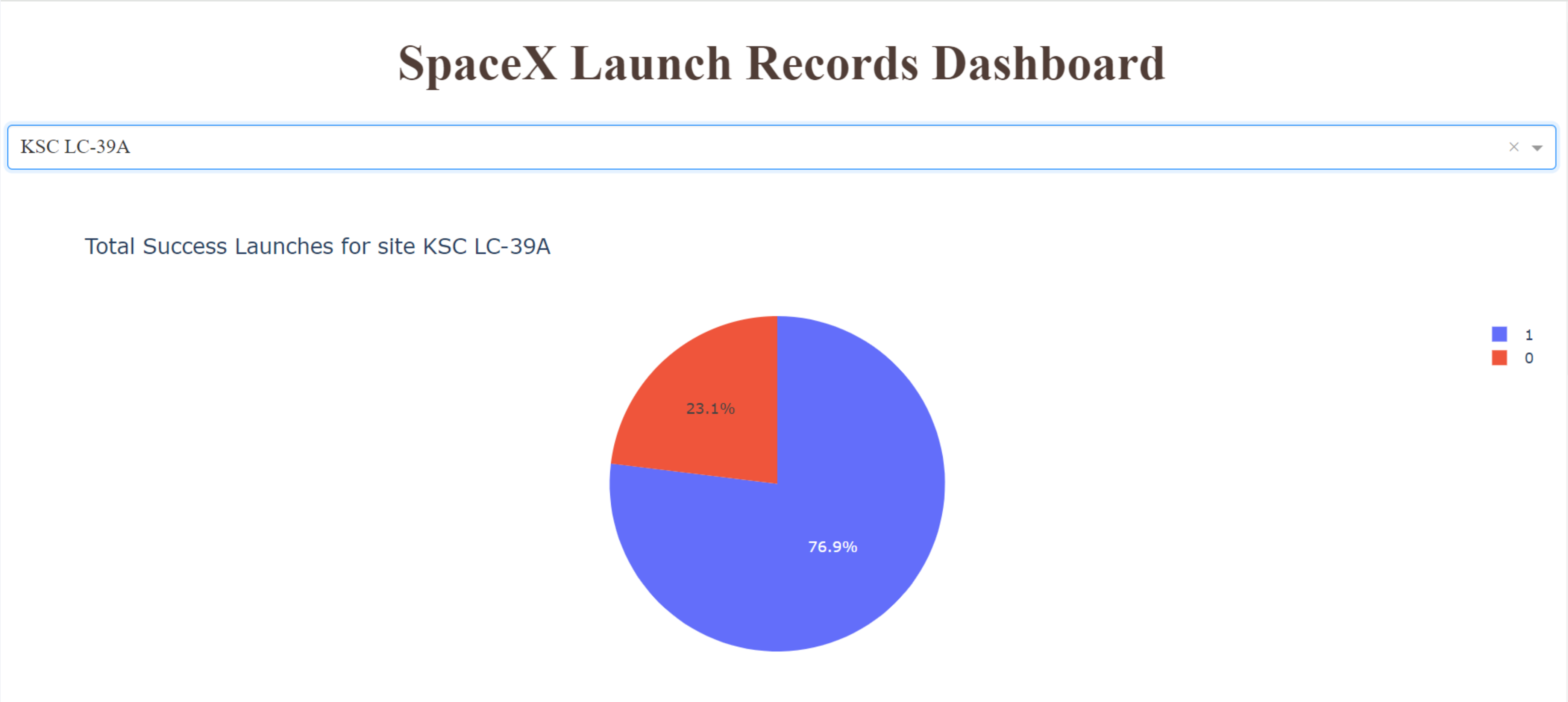
Success Launch Outcome for all sites

- KSC LC-39A has most success launch outcomes compare to other three sites



The highest Launch Outcome success ratio site

- KSC LC-39A also has the highest success launch ration compare to other three sites

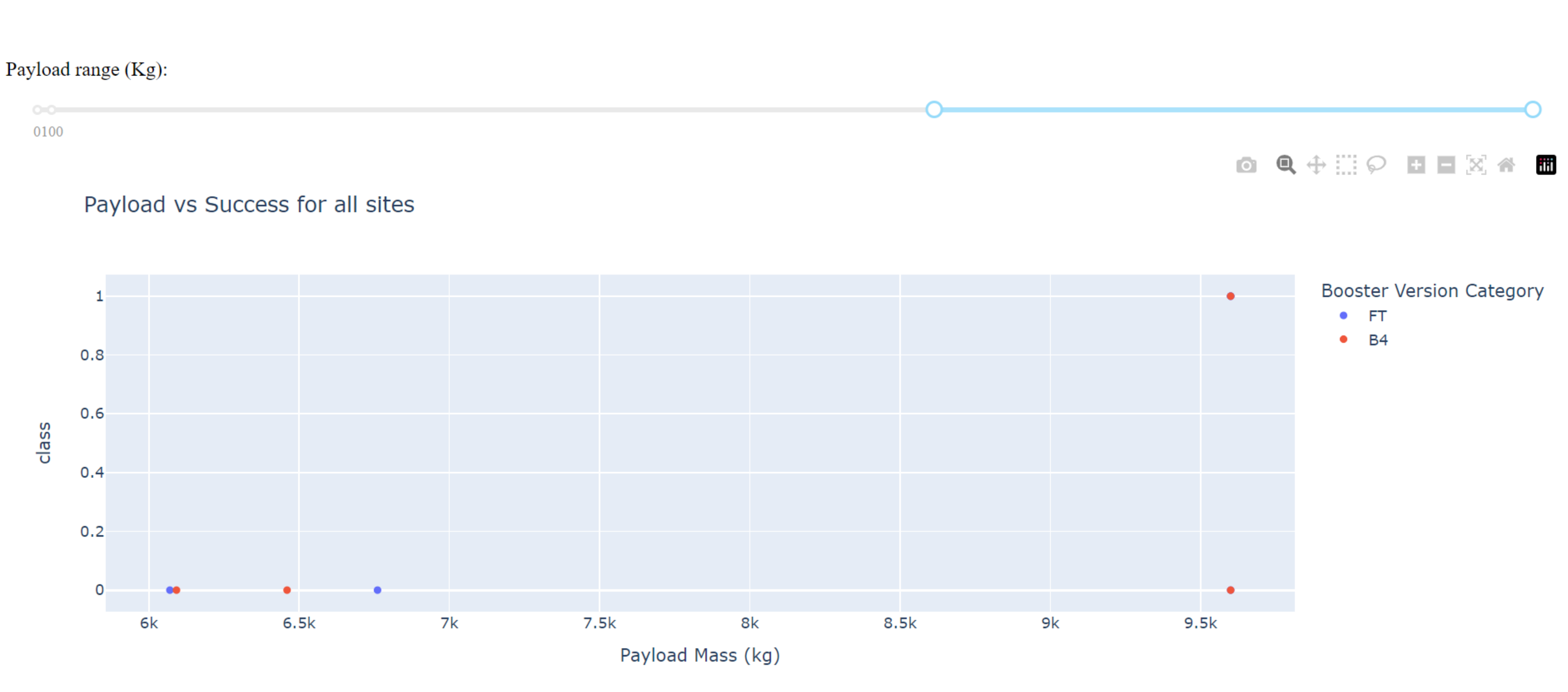


Payload mass vs. Launch Outcome for all sites (1)

- Under 6,000 kg, FT booster has best success outcome

Payload mass vs. Launch Outcome for all sites (2)

- Between 6,000 kg and 10,000 kg, only one launch is successful



Section 5

Predictive Analysis (Classification)

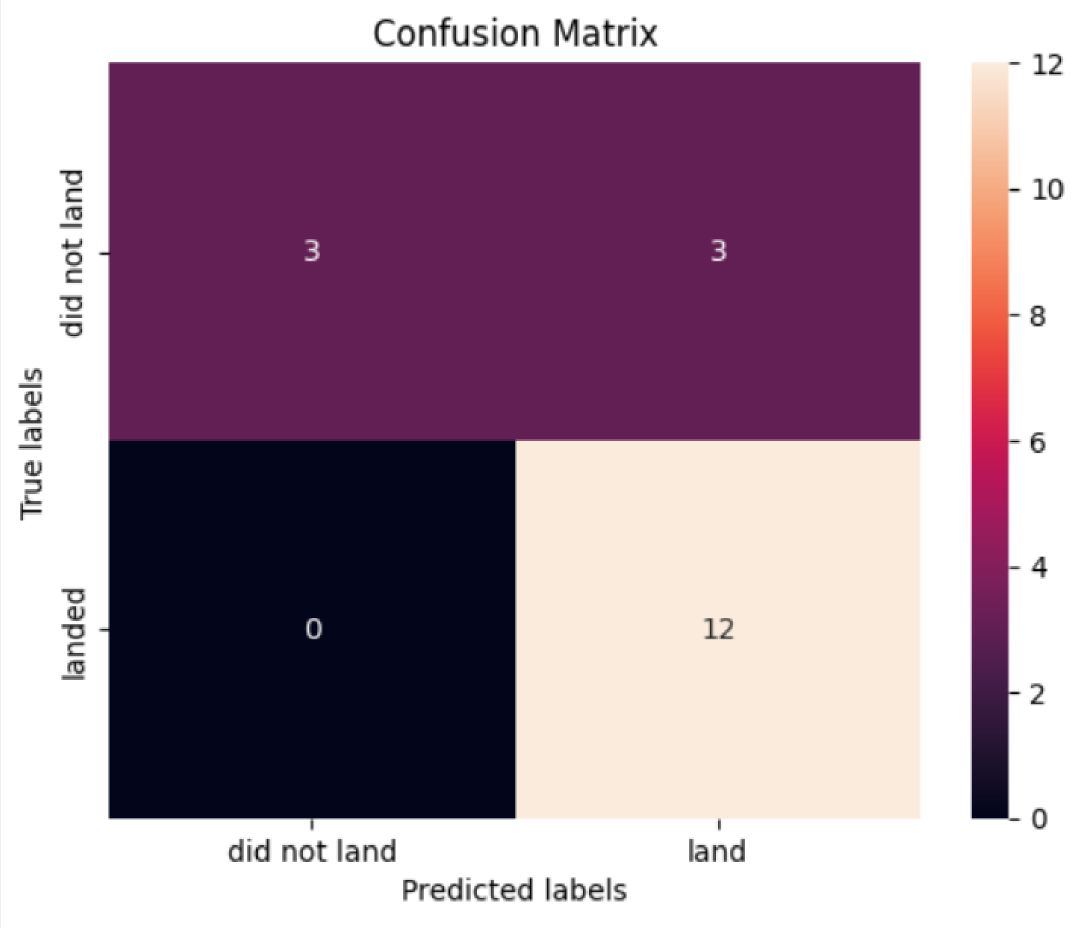
Classification Accuracy

- From the table, we can see that all four methods has the same test accuracy
- So we can not pick the best, since all methods have the same test accuracy

	Accuracy	Test Accuracy
Log Regression	0.846	0.833
SVM	0.848	0.833
Decision Tree	0.863	0.833
KNN	0.848	0.833

Confusion Matrix

- Since all four methods have the same test score accuracy, they have the same confusion matrix
- All the methods have the same problem with false positive



Conclusions

- Landing outcome success rate has improvement trend from 2014 to 2020
- Launch site KSC LC-39A has the best success rate compare to other three sites
- ES-L1, GEO, HEO, and SSO orbit launch has 100% success rate
- Machine Learning Predictive model has the same test accuracy, any of the four models can be used
- The distance from launch site to its proximities need to have a certain distance

Thank you!